Passive and Pervasive Use of a Bilingual Dictionary in Statistical Machine Translation

Liling Tan[1], **Josef van Genabith**[1] **and Francis Bond**[2]
Universität des Saarland[1] / Campus, Saarbrücken, Germany
Nanyang Technological University[2] / 14 Nanyang Drive, Singapore
liling.tan@uni-saarland.de, josef.van_genabith@dfki.de,
bond@ieee.org

Abstract

There are two primary approaches to the use bilingual dictionary in statistical machine translation: (i) the passive approach of appending the parallel training data with a bilingual dictionary and (ii) the pervasive approach of enforcing translation as per the dictionary entries when decoding. Previous studies have shown that both approaches provide external lexical knowledge to statistical machine translation thus improving translation quality. We empirically investigate the effects of both approaches on the same dataset and provide further insights on how lexical information can be reinforced in statistical machine translation.

1 Introduction

Statistical Machine Translation (SMT) obtains the best translation, e_{best}, by maximizing the conditional probability of the foreign sentence given the source sentence, p(f|e), and the a priori probability of the translation, p_{LM}(e) (Brown, 1993).

$$\begin{aligned} e_{best} &= \underset{e}{\operatorname{argmax}} \quad p(e|f) \\ &= \underset{e}{\operatorname{argmax}} \quad p(f|e) \quad p_{LM}(e) \end{aligned}$$

State-of-art SMT systems rely on (i) large bilingual corpora to train the translation model p(f|e) and (ii) monolingual corpora to build the language model, p_{LM}(e).

One approach to improve the translation model is to extend the parallel data with a bilingual dictionary prior to training the model. The primary motivation to use additional lexical information for domain adaptation to overcome the out-of-vocabulary words during decoding (Koehn and Schroeder, 2007; Meng et al. 2014; Wu et al. 2008). Alternatively, adding in-domain lexicon to parallel data has also shown to improve SMT. The intuition is that by adding extra counts of bilingual lexical entries, the word alignment accuracy improves, resulting in a better translation model (Skadins et al. 2013; Tan and Pal, 2014; Tan and Bond, 2014).

Another approach to use a bilingual dictionary is to hijack the decoding process and force word/phrase translations as per the dictionary entries. Previous researches used this approach to explore various improvements in industrial and academic translation experiments. For instance, Tezcan and Vandeghinste (2011) injected a bilingual dictionary in the SMT decoding process and integrated it with Computer Assisted Translation (CAT) environment to translate documents in the technical domain. They showed that using a dictionary in decoding improves machine translation output and reduces post-editing time of human translators. Carpuat (2009) experimented with translating sentences in discourse context by using a discourse specific dictionary annotations to resolve lexical ambiguities and showed that this can potentially improve translation quality.

In this paper, we investigate the improvements made by both approaches to use a bilingual dictionary in SMT. We refer to the first approach of extending the parallel data with dictionary as the ***passive*** use and the latter approach of hijacking the decoding process as the ***pervasive*** use of dictionary in statistical machine translation.

Different from the normal use of a dictionary for the purpose of domain adaptation where normally, a domain-specific lexicon is appended to a translation model trained on generic texts, we are investigating the use of an in-domain dictionary in statistical machine translation.

More specifically, we seek to understand how much improvement can be made by skewing the lexical information towards the passive and pervasive use of the dictionary in statistical machine

Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), pages 30–34,
Beijing, China, July 31, 2015.

translation.

2 Passive vs Pervasive Use of Dictionary

We view both the passive and the pervasive use of a dictionary in statistical machine translation as a type of ***lexically constrained statistical hybrid MT*** where in the passive use, the dictionary acts a a supplementary set of bi-lexical rules affecting word and phrase alignments and the resulting translation model and in the pervasive use, the dictionary constraints the decoding search space enforcing translations as per the dictionary entries.

To examine the ***passive use*** of a dictionary, we explore the effects of adding the lexicon *n* number of times to the training data until the performance of the machine translation degrades.

For the ***pervasive use*** of a dictionary, we assign a uniform translation probability to possible translations of the source phrase. For instance, according to the dictionary, the English term "abnormal hemoglobin" could be translated to 異常ヘモグロビン or 異常血色素, we assign the translation probability of 0.5 to both Japanese translations, i.e. p(異常ヘモグロビン | *abnormal hemoglobin*) = p(異常血色素 | *abnormal hemoglobin*) = 0.5. If there is only one translation for a term in the dictionary, we force a translation from the dictionary by assigning the translation probability 1.0 to the translation.

One issue with the pervasive use of dictionary translations is the problem of compound phrases in the test sentence that are made up of ***component phrases*** in the dictionary. For instance, when decoding the sentence, "*Here was developed a phase shift magnetic sensor system composed of two sets of coils , amplifiers , and phase shifts for sensing and output* .", we fetch the following entries from the dictionary to translate the underlined multiword term:

- *magnetic* =磁気
- *sensor* = センサ, センサー, 感知器, 感知部, 感応素子, 検出変換器, 変換素子, 受感部, 感覚器, センサー
- *system* = 組織体制, 制度, 子系, 系列, システム, 体系, 方式, 系統, 秩序, 体制, 組織, 一方式
- *magnetic sensor* = 磁気センサ
- *sensor system* = センサシステム, センサ系, センサーシステム

In such a situation, where the dictionary does not provide a translation for the complete multiword string, we set the preference for the dictionary entry with the longest length in the direction from left to right and select "*magnetic sensor*" + "*system*" entries for forced translation.[1]

Finally, we investigate the effects of using the bilingual dictionary both passively and pervasively by appending the dictionary before training and hijacking the decoding by forcing translations using the same dictionary.

3 Experimental Setup

We experimented the passive and pervasive uses of dictionary in SMT using the Japanese-English dataset provided in the Workshop for Asian Translation (Toshiaki et al. 2014). We used the Asian Scientific Paper Excerpt Corpus (ASPEC) as the training corpus used in the experiments. The ASPEC corpus consists of 3 million parallel sentences extracted from Japanese-English scientific abstracts from Japan's Largest Electronic Journal Platform for Academic Societies (J-STAGE). In our experiments we follow the setup of the WAT shared task with 1800 development and test sentences each from the ASPEC corpus.

We use the Japanese-English (JA-EN) translation dictionaries (JICST, 2004) from the Japan Science and Technology Corporation. It contains 800,000 entries[2] for technical terms extracted from scientific and technological documents. Both the parallel data and the bilingual dictionary are tokenized with the `MeCab` segmenter (Kudo et al. 2004).

Dataset	Japanese	English
`Train`	86M	78M
`Dev.`	47K	44K
`Test`	47K	44K
`Dict.`	2.1M	1.7M

Table 1: Size of Training (`Train`), Development (`Dev.`) and Test (`Test`) Dataset from the ASPEC Corpus and JICST Dictionary (`Dict.`).

Table 1 presents the number of tokens in the ASPEC corpus and the JICST dictionary. On average 3-4 dictionary entries are found for each sentence

[1] Code to automatically convert sentences into XML-input with pervasive dictionary translations for the Moses toolkit is available at `http://tinyurl.com/pervasive-py`.

[2] 2.1M JA and 1.7M EN tokens

in the WAT development set.

For all experiments we used the phrase-based SMT implemented in the Moses toolkit (Koehn et al, 2007) with the following experimental settings:

- `MGIZA++` implementation of IBM word alignment model 4 with grow-diagonal-final-and heuristics for word alignment and phrase-extraction (Och and Ney, 2003; Koehn et al., 2003; Gao and Vogel, 2008)
- Bi-directional lexicalized reordering model that considers monotone, swap and discontinuous orientations (Koehn et al., 2005 and Galley and Manning, 2008)
- Language modeling is trained using `KenLM` with maximum phrase length of 5 with Kneser-Ney smoothing (Heafield, 2011; Kneser and Ney, 1995)
- Minimum Error Rate Training (MERT) (Och, 2003) to tune the decoding parameters.
- For English translations, we trained a true-casing model to keep/reduce tokens' capitalization to their statistical canonical form (Wang et al., 2006; Lita et al., 2003) and we recased the translation output after the decoding process

Addtionally, we applied the following methods to optimize the phrase-based translation model for efficiency:

- To reduce the size of the language model and the speed of querying the model when decoding, we used the binarized trie-based quantized language model provided in `KenLM` (Heafield et al. 2013, Whittaker and Raj, 2001)
- To minimize the computing load on the translation model, we compressed the phrase-table and lexical reordering model using the `cmph` tool (Junczys-Dowmunt, 2012)

For the passive use of the dictionary, we simply appended the dictionary to the training data before the alignment and training process. For the pervasive use of the dictionary, we used the `xml-input` function in the Moses toolkit to force lexical knowledge in the decoding process[3].

[3]http://www.statmt.org/moses/?n=Advanced.Hybrid#ntoc1

4 Results

Table 2 presents the BLEU scores of the Japanese to English (JA-EN) translation outputs from the phrase-based SMT system on the WAT test set. The leftmost columns indicate the number of times a dictionary is appended to the parallel training data (*Baseline* = 0 times, *Passive x1* = 1 time). The rightmost columns present the results from both the passive and pervasive use of dictionary translations, with exception to the top-right cell which shows the baseline result of the pervasive dictionary usage without appending any dictionary.

	- Pervasive	+ Pervasive
Baseline	16.75	16.87
Passive x1	16.83	17.30**
Passive x2	**17.31****	16.87
Passive x3	17.26*	17.06
Passive x4	17.14*	**17.38****
Passive x5	16.82	17.29**

Table 2: BLEU Scores for Passive and Pervasive Use of the Dictionary in SMT (Japanese to English)

By repeatedly appending the dictionary to the parallel data, the BLEU scores significantly[4] improves from 16.75 to 17.31. Although the system's performance degrades when adding the dictionary passively thrice, the score remains significantly better than baseline. The pervasive use of the dictionary improves the baseline without the passive of the dictionary. The best performance is achieved when the dictionary is passively added four times with the pervasive use of the dictionary during decoding.

The fluctuations in improvement from coupling the passive and pervasive use of an in-domain dictionary give no indication of how both approaches should be used in tandem. However, using either or both the approaches improves the translation quality of the baseline system.

Table 3 presents the BLEU scores of the English to Japanese (EN-JA) translation outputs from the phrase-based SMT system on the WAT test set. Similarly, the passive use of dictionary outperforms the baseline but the pervasive use of dictionary consistently reported worse BLEU scores significantly.

Different from the JA-EN translation the pervasive use of dictionary consistently performs worse

[4]*: p-value$<$0.1, **: p-value$<$0.001

	- Pervasive	+ Pervasive
Baseline	23.91	23.14**
Passive +1	24.12*	23.13**
Passive +2	23.79	22.86**
Passive +3	**24.14***	**23.29****
Passive +4	24.13*	23.16**
Passive +5	23.67	22.71**

Table 3: BLEU Scores for Passive and Pervasive Use of Dictionary in SMT (English to Japanese)

than the baseline. Upon random manual checking of the MT output, there are many instances where the technical/scientific term in the dictionary is translated correctly with only the passive use of the dictionary. However, it unclear whether the overall quality of the translations have degraded from the pervasive use of the dictionary given the slight, though significant, decrease in BLEU scores.

5 Conclusion

Empirically, both passive and pervasive use of a in-domain dictionary to extend statistical machine translation models with lexical knowledge modestly improve translation quality.

Interestingly, the fact that adding the in-domain dictionary information multiple times to the training data improves MT suggests that there may be a critical probability mass that a lexicon can impact the word and phrasal alignments in a corpus. This may provide insight on optimizing the weights of the salient in-domain phrases in the phrase table.

Although the pervasive use of dictionary information provides minimal or no improvements to the BLEU scores in our experiments, it remains relevant in industrial machine translation where terminological standardization is crucial in ensuring consistent translations of technical manuals or legal texts where incorrect use of terminology may have legal consequences (Porsiel, 2011).

The reported BLEU improvements from the passive information use of dictionary are good indication of improved machine translation quality but BLEU scores deterioration in the pervasive use only indicates that the output is not the same as the reference translation. Further manual evaluation is necessary to verify the poor performance of the pervasive use of dictionary information in machine translation.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27.

Michel Galley and Christopher D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria, August.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

JICST, editor. 2004. *JICST Japanese-English translation dictionaries*. Japan Information Center of Science and Technology.

Marcin Junczys-Dowmunt. 2012. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *The Prague Bulletin of Mathematical Linguistics*, 98:63–74.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North*

American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 48–54.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit, vol. 5, pp. 79-86.*

Lucian Vlad Lita, Abraham Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. tRuEcasIng. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

Fandong Meng, Deyi Xiong, Wenbin Jiang, and Qun Liu. 2014. Modeling term translation for document-informed machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 546–556, Doha, Qatar, October.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st workshop on asian translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Raivis Skadiņš, Mārcis Pinnis, Tatiana Gornostay, and Andrejs Vasiļjevs. 2013. Application of online terminology services in statistical machine translation. pages 281–286.

Liling Tan and Francis Bond. 2014. Manipulating input data in machine translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201–206, Baltimore, Maryland, USA, June.

Arda Tezcan and Vincent Vandeghinste. 2011. Smtcat integration in a technical domain: Handling xml markup using pre & post-processing methods. *Proceedings of EAMT 2011*.

Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Stephan Vogel and Christian Monson. 2004. Augmenting manual dictionaries for statistical machine translation systems. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Wei Wang, Kevin Knight, and Daniel Marcu. 2006. Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 1–8, New York City, USA, June. Association for Computational Linguistics.

Edward WD Whittaker and Bhiksha Raj. 2001. Quantization-based language model compression. In *Proceedings of INTERSPEECH*, pages 33–36.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 993–1000.